The Gene Ontology Task at BioCreative IV

Yuqing Mao[1], Kimberly Van Auken[2], Donghui Li[3], Cecilia N. Arighi[4], Zhiyong Lu[1,*]

[1]National Center for Biotechnology Information (NCBI), 8600 Rockville Pike, MD 20817 USA [2]WormBase, Division of Biology, California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125 USA [3]Department of Plant Biology, The Arabidopsis Information Resource, Carnegie Institution for Science, Stanford, CA 94305, USA [4]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19711, USA

[*]Corresponding author: Tel: 301 594 7089, E-mail: Zhiyong.Lu@nih.gov

Abstract

Gene Ontology (GO) annotation is a common task among model organism database (MOD) groups. It is a very time-consuming and labor-intensive task, thus often considered as one of the bottlenecks in literature curation. There is a growing need for semi- or fully-automated GO curation techniques that will help database curators rapidly and accurately identify gene function information in full-length articles. Despite multiple attempts in the past, few studies have proven to be useful with regard to assisting real-world GO curation. The lack of relevant training data and opportunities for interaction between text mining developers and GO curators has limited the advances in algorithm development and corresponding use in practical circumstances. To this end, we organized a text-mining challenge task for literature-based GO annotation in BioCreative IV. More specifically, we developed two sub-tasks: a) to automatically locate text passages that contain GO-relevant information (a text retrieval task) and b) to automatically identify relevant GO terms for the genes in a given article (a concept recognition task). With the support from five MODs, we provided teams with nearly 4,000 unique text passages that served as the basis for each GO annotation in our task data. Such evidence text information has long been recognized as critical for text-mining algorithm development but was never made available due to the high cost of curation. In total, seven teams participated in the challenge task. From the team results, we find an overall improvement in performance for recognizing GO terms when comparing to similar task results in the past. Future work should focus on improving performance of GO concept recognition and incorporating practical benefits of text-mining tools into real-world GO annotation.

Introduction

Manual Gene Ontology (GO) annotation is the task of human curators assigning gene functional information using GO terms through reading the biomedical literature. This is a common task among Model Organism Database (MOD) groups (1) and can be time-consuming and labor-intensive. Thus, manual GO annotation is often considered one of the bottlenecks in literature-based biocuration (2). As a result, many MODs can only afford to curate a fraction of relevant articles. For instance, the curation team of The Arabidopsis Information Resource (TAIR) has been able to curate less than 30% of newly published articles that contain information about Arabidopsis genes (3).

Recently, there is a growing interest for building automatic text-mining tools to assist manual biocuration (4-10), including systems that aim to help database curators rapidly and accurately identify gene function information in full-length articles (11,12). Although automatically mining GO terms from full-text articles is not a new problem in BioNLP, few studies have proven to be useful with regard to assisting real-world GO curation. The lack of access to relevant evidence text associated with GO annotations and limited opportunities for interaction with actual GO curators have been recognized as the major difficulties in algorithm development and corresponding application in practical circumstances (12,13). As such, in BioCreative IV, not only do we plan to provide teams with article-level gold-standard GO annotations for each full-text article as has been done in the past, but we will also provide evidence text for each GO annotation with the help from expert GO curators. That is, to best help text-mining tool advancement, evidence text passages that support each GO annotation will be provided in addition to the usual GO annotations which typically include three distinct elements: gene or gene product, GO term, and GO evidence code.

Also as we know from past BioCreative tasks, recognizing gene names and experimental codes from full text are difficult tasks on their own (14-17). Hence, to encourage teams to focus on GO term extraction, we proposed, for this task, to separate gene recognition from GO term and evidence code selection by including both the gene names and associated NCBI Gene identifiers in the task data sets.

Specifically, we propose two challenge tasks towards automated GO concept recognition from full-length articles:

Task A: Retrieving GO evidence text for relevant genes

GO evidence text is critical for human curators to make associated GO annotations. For a given GO annotation, multiple evidence passages may appear in the paper, some being more specific with experimental information while others may be more succinct about the gene function. For this sub-task, participants are given as input full-text articles together with relevant gene information. For system output, teams have to submit a list of GO evidence sentences for each of the input genes in the paper. Manually curated GO evidence passages will be used as the gold standard for evaluating team submissions. Each team is allowed to submit 3 runs.

Task B: Predicting GO terms for relevant genes

This sub-task is a step towards the ultimate goal of using computers for assisting human GO curation. As in Task A, participants are given as input full text articles with relevant gene information. For system output, teams are asked to return a list of relevant GO terms for each of the input genes in a paper. Manually curated GO annotations will be used as the gold standard for evaluating team predictions. Similar to Task A, each team is allowed to submit 3 runs.

Generally speaking, the first sub-task is a text retrieval task while the second can be seen as a multi-class text classification problem where each GO term represents a distinct class label. In the BioNLP research domain, the first sub-task is similar to the BioCreative I GO sub-task 2.1 (12), BioCreative II Interaction Sentence sub-task (14), and automatic GeneRIF identification (18-20). The second sub-task is similar to the BioCreative I GO sub-task 2.2 (12) and is also closely related to the problem of semantic indexing of biomedical literature such as automatic indexing of biomedical publications with MeSH terms (21-24).

Methods

Corpus annotation

A total of 8 professional GO curators from five different MODs (FlyBase; MaizeGDB; RGD; TAIR; WormBase) contributed to the development of the task data. To create the annotated corpus, each curator was asked, in addition to their routine annotation of gene-related GO information, to mark up the associated evidence text in each paper that supports those annotations using a Web-based annotation tool. To provide complete data for text-mining system development (i.e., both positive and negative training data), curators were asked to select evidence text exhaustively throughout the paper (25).

For obtaining high-quality and consistent annotations across curators, detailed annotation guidelines were developed and provided to the curators. In addition, each curator was asked to practice on a test document following the guidelines before they begin curating task documents. Due to the significant workload and limited number of curators per group, each paper was only annotated by a single curator.

Evaluation measures

For Task A evaluation, traditional precision (P), recall (R) and F_1 score (F_1) are reported when comparing the submitted gene-specific sentence list against the gold standard. We computed the numbers of true positives (TP) and false positives (FP) in two ways: the first one (exact match) is a strict measure that requires the returned sentences exactly match the sentence boundary of human markups while the second (overlap) is a more relaxed measure where a prediction is considered correct (i.e. TP) as long as the submitted sentence overlaps with the gold standard.

$$P = \frac{tp}{tp+fp},\ R = \frac{tp}{tp+fn},\ F_1 = 2\cdot\frac{P\times R}{P+R}$$

For the Task B evaluation, gene-specific GO annotations in the submissions will be compared with the gold standard. In addition to the traditional precision, recall and F1 score, hierarchical Precision (hP), Recall (hR) and F-score (hF_1) will also be computed where common ancestors in both the computer-predicted and human-annotated GO terms are considered. The second set of measures were proposed to reflect the hierarchical nature of GO: a gene annotated with one GO term is implicitly annotated with all of the term's parents, up to the root concept (26,27). Such a measure takes into account that "predictions that are close to the oracle label should score better than predictions that are in an unrelated part of the hierarchy." (26) Specifically, the hierarchical measures are computed as:

$$hP = \frac{\sum_i \left|\hat{G}_i \cap \hat{G}'_i\right|}{\sum_i \left|\hat{G}'_i\right|},\ hR = \frac{\sum_i \left|\hat{G}_i \cap \hat{G}'_i\right|}{\sum_i \left|\hat{G}_i\right|},\ hF_1 = 2\cdot\frac{hP\cdot hR}{hP+hR}$$

$$\hat{G}_i = \{\bigcup_{G_k \in G_i} Ancestors(G_k)\}$$

$$\hat{G}'_i = \{\bigcup_{G'_k \in G'_i} Ancestors(G'_k)\}$$

where $\hat{G}_i$ and $\hat{G}'_i$ are the respective sets of ancestors of the computer-predicted and human-annotated GO terms for the *i*th gene.

Results

The BC4GO corpus

The task participants were provided with three data sets comprising a total of 200 full-text articles in the BioC XML format (28). Our evaluation for the two sub-tasks was to respectively assess teams' ability to return relevant sentences and GO terms for each given gene in the 50 test articles. Hence, we show in Table 1 the overall statistics of the BC4GO corpus including the numbers of genes, gene-associated GO terms and evidence text passages. For instance, in the 50 test articles, 194 genes were associated with 644 GO Terms, and 1,681 evidence text passages, respectively. We refer interested readers to (25) for a detailed description of the BC4GO corpus.

Table 1. Overall statistics of the BC4GO corpus.

Curated Data	Training Set	Dev. Set	Test Set
Full text articles	100	50	50
Genes in those articles	300	171	194
Gene-associated **passages** in those articles	2,234	1,247	1,681
Gene-associated **GO terms** in those articles	954	575	644

Team participation results

Overall, seven teams (3 from Americas, 3 from Asia, and 1 from Europe) participated in the GO task. In total, they submitted 32 runs: 15 runs from five different teams for Task A, and 17 runs from six teams for Task B.

Team Results of Task A

Table 2 shows the results of 15 runs submitted by the five participating teams in Task A. Run 3 from Team 238 achieved the highest F_1 score in both exact match (0.270) and overlap (0.387) calculations. Team 238 is also the only team who submitted results for all 194 genes from the input of the test set. The highest recall is 0.424 in exact match and 0.716 in overlap calculations by the same run (Team 264, run 1), respectively. The highest precision is 0.220 in exact match by Team 238 Run 2 and 0.354 in overlap by Team 183, Run 2.

Table 2. Team results for Task A using traditional Precison (P), Recall (R) and F-Mesuare (F1). Both strict exact match and relaxed overlap measure are considered.

Team	Run	Genes	Passages	Exact match			Overlap		
				P	R	F_1	P	R	F_1
183	1	173	1,042	0.206	0.128	0.158	0.344	0.213	0.263
183	2	173	1,042	0.217	0.134	0.166	**0.354**	0.220	0.271
183	3	173	1,042	0.107	0.066	0.082	0.204	0.127	0.156
237	1	23	54	0.185	0.006	0.012	0.333	0.011	0.021
237	2	96	2,755	0.103	0.171	0.129	0.214	0.351	0.266
237	3	171	3,717	0.138	0.305	0.190	0.213	0.471	0.293
238	1	194	2,698	0.219	0.352	**0.270**	0.313	0.503	0.386
238	2	194	2,362	**0.220**	0.310	0.257	0.314	0.442	0.367
238	3	194	2,866	0.214	0.366	**0.270**	0.307	0.524	**0.387**
250	1	161	3,297	0.146	0.286	0.193	0.239	0.469	0.317
250	2	140	2,848	0.153	0.259	0.193	0.258	0.437	0.325
250	3	161	3,733	0.140	0.311	0.193	0.226	0.503	0.312
264	1	167	13,533	0.052	**0.424**	0.093	0.088	**0.716**	0.157
264	2	111	2,243	0.037	0.049	0.042	0.077	0.103	0.088
264	3	111	2,241	0.037	0.049	0.042	0.077	0.103	0.088

Team Results of Task B

Table 3 shows the results of 17 runs submitted by the six participating teams in Task B. Run 1 from Team 183 achieved the highest F_1 score in traditional (0.134) and hierarchical measures (0.338). The same run also obtained the highest precision of 0.117 in exact match while the highest precision in hierarchical match is 0.415 obtained by the Run 1 of Team 237. However, note that this run only returned 37 GO terms for 23 genes. The highest recall is 0.306 and 0.647 in the two measures by Run 3 of Team 183.

Table 3. Team results for the Task B using traditional Precision (P), Recall (R) and F1-measure (F1) and hierarchical precision (hP), recall (hR) and F1-measure (hF1).

Team	Run	Genes	GO terms	Exact match			Hierarchical match		
				P	R	F_1	hP	hR	hF_1
183	1	172	860	**0.117**	0.157	**0.134**	0.322	0.356	**0.338**
183	2	172	1720	0.092	0.245	**0.134**	0.247	0.513	0.334
183	3	172	3440	0.057	**0.306**	0.096	0.178	**0.647**	0.280
220	1	50	2639	0.018	0.075	0.029	0.064	0.190	0.096
220	2	46	1747	0.024	0.065	0.035	0.087	0.158	0.112
237	1	23	37	0.108	0.006	0.012	**0.415**	0.020	0.039
237	2	96	2424	0.108	0.068	0.029	0.084	0.336	0.135
237	3	171	4631	0.037	0.264	0.064	0.150	0.588	0.240
238	1	194	1792	0.054	0.149	0.079	0.243	0.459	0.318
238	2	194	555	0.088	0.076	0.082	0.250	0.263	0.256
238	3	194	850	0.029	0.039	0.033	0.196	0.310	0.240
243	1	109	510	0.073	0.057	0.064	0.249	0.269	0.259
243	2	104	393	0.084	0.051	0.064	0.280	0.248	0.263
243	3	144	2538	0.030	0.116	0.047	0.130	0.477	0.204
250	1	171	1389	0.052	0.112	0.071	0.174	0.328	0.227
250	2	166	1893	0.049	0.143	0.073	0.128	0.374	0.191
250	3	132	453	0.095	0.067	0.078	0.284	0.161	0.206

Discussion and Conclusions

As mentioned earlier, our task is related to a few previous challenge tasks on biomedical text retrieval and semantic indexing. In particular, our task resembles the earlier GO task in BioCreative I (12). On the other hand, our two sub-tasks are also different from the previous tasks. For the passage retrieval task, we only provide teams with pairs of <gene, document> and ask their systems to return relevant evidence text while <gene, document, GO terms> triples were provided in the earlier task.

For the GO term prediction task, we provided teams with the same <gene, document> pairs and asked their systems to return relevant GO terms. In addition to such input pairs, the expected number of GO terms and their associated GO branches (molecular function, biological process, and cellular component) returned were also provided in the earlier task. Another difference is that along with each predicted GO term for the given gene in the given document, output of associated evidence text is also required in the earlier task.

Finally, the evaluation mechanism differed in the two challenge events. We provided the reference data prior to the team submission and preformed standard evaluation. By contrast, in the BioCreative I GO task, no gold-standard evaluation data were provided before the team submission. Instead, expert GO curators were asked to manually judge the team submitted results. Such a post-hoc analysis could miss true positives not returned by teams and would not permit evaluation of new systems after the challenge.

In summary, we provided less input information to teams in both sub-tasks and followed protocols of standard challenge evaluation – two major differences between our task and the previous BioCreative I task (12). This is partly because we aim to have our tasks resemble real-world GO annotation more closely, where the only input to human curators is the set of documents. Despite these differences, we were intrigued by any potential improvement in the task results due to the advancement of text mining research in recent years. Since the ultimate goal of the task is to find GO terms, the results of Task B are of more interest and significance in this aspect, though evidence sentences are of course important for reaching this goal. By comparing the team results in the two challenge events (Table 3 above vs. Table 5 in (12)), we can observe a general trend of performance increase on this task over time. For example, the best-performing team in 2005 (12) was only able to predict 78 TPs (out of 1227 in gold standard) – a recall of less than 7% – while there are several teams in our task who obtained recall values between 10% and 30%. The numbers are even greater when measured by taking account of the hierarchical nature of the Gene Ontology.

Despite these encouraging results, overall team results suggest that automatically mining GO terms from literature remains very challenging due to difficulties in multiple aspects: First, the number of GO terms (class labels for classification) is extremely large: there are over 40,000 unique GO concepts to date. Second, GO terms (and associated synonyms) are designed for unifying gene function annotations rather than for text mining, and are therefore rarely found verbatim in the article. For example, our analysis shows that only about 1/3 of the annotated GO terms in our corpus can be found using exact matches in their corresponding articles. On the other hand, not every match related to a GO concept is annotated. Instead, only those GO terms that represent experimental findings in a given full-text paper are selected. Hence, automatic methods must be able to filter irrelevant mentions that share names with GO terms (e.g. the GO term ‘growth’ is a common word in articles, but additional contextual information would be required to determine if this relatively high-level term should be used for GO annotation purposes). Finally, human annotation data for building statistical/machine-learning approaches is still lacking. Despite our best efforts, we are only able to include 200 annotated articles in our corpus, which contains evidence text for only 1,311 GO terms.

Our challenge task was inspired and developed in response to the actual needs of GO manual annotation. However, compared to the real-world GO annotation, the BioCreative challenge task is simplified in two aspects: a) gene information is provided to the teams while in reality they are unknown; and b) extraction of GO evidence code information is not required for our task while it is an essential part of the GO annotations in practice. Further investigation of automatic extraction of gene and evidence code information, along with corresponding GO terms, remains as future work.

Acknowledgments

We are grateful to the support of our corpus curators. We also thank Lynette Hirschman, John Wilbur, Cathy Wu, Kevin Cohen, Martin Krallinger, and Thomas Wiegers from the BioCreative IV organizing committee for their support, and Judith Blake, Andrew Chatr-aryamontri, Sherri Matis, Fiona McCarthy, Sandra Orchard, and Phoebe Roberts from the BioCreative IV User Advisory Group for their helpful discussions. This research is supported by NIH Intramural Research Program, National Library of Medicine (YM & ZL).

References

1. Balakrishnan, R., Harris, M.A., Huntley, R.*, et al.* (2013) A guide to best practices for Gene Ontology (GO) manual annotation. *Database : the journal of biological databases and curation*, **2013**, bat054.
2. Lu, Z., Hirschman, L. (2012) Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database : the journal of biological databases and curation*, **2012**, bas043.
3. Li, D., Berardini, T.Z., Muller, R.J.*, et al.* (2012) Building an efficient curation workflow for the Arabidopsis literature corpus. *Database : the journal of biological databases and curation*, **2012**, bas047.
4. Wu, C.H., Arighi, C.N., Cohen, K.B.*, et al.* (2012) BioCreative-2012 virtual issue. *Database : the journal of biological databases and curation*, **2012**, bas049.
5. Arighi, C.N., Carterette, B., Cohen, K.B.*, et al.* (2013) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database : the journal of biological databases and curation*, **2013**, bas056.
6. Wei, C.H., Harris, B.R., Li, D.*, et al.* (2012) Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database : the journal of biological databases and curation*, **2012**, bas041.
7. Neveol, A., Wilbur, W.J., Lu, Z. (2012) Improving links between literature and biological data with text mining: a case study with GEO, PDB and MEDLINE. *Database : the journal of biological databases and curation*, **2012**, bas026.
8. Wei, C.-H., Kao, H.-Y., Lu, Z. (2012) PubTator: A PubMed-like interactive curation system for document triage and literature curation. *Proceedings of the BioCreative 2012 workshop*, Washington, D.C., pp. 20-24.
9. Wei, C.-H., Kao, H.-Y., Lu, Z. (2013) PubTator: a Web-based text mining tool for assisting Biocuration. *Nucleic Acids Res*, **41**, W518-W522.
10. Neveol, A., Wilbur, W.J., Lu, Z. (2011) Extraction of data deposition statements from the literature: a method for automatically tracking research results. *Bioinformatics*, **27**, 3306-3312.
11. Van Auken, K., Jaffery, J., Chan, J.*, et al.* (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*, **10**, 228.

12. Blaschke, C., Leon, E.A., Krallinger, M.*, et al.* (2005) Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, **6 Suppl 1**, S16.
13. Camon, E.B., Barrell, D.G., Dimmer, E.C.*, et al.* (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, **6 Suppl 1**, S17.
14. Krallinger, M., Leitner, F., Rodriguez-Penagos, C.*, et al.* (2008) Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol*, **9 Suppl 2**, S4.
15. Lu, Z., Kao, H.Y., Wei, C.H.*, et al.* (2011) The gene normalization task in BioCreative III. *BMC Bioinformatics*, **12 Suppl 8**, S2.
16. Lu, Z., Wilbur, W.J. (2010) Overview of BioCreative III Gene Normalization. *Proceedings of the BioCreative III workshop*, Bethesda, USA, pp. 24-45.
17. Van Landeghem, S., Bjorne, J., Wei, C.H.*, et al.* (2013) Large-scale event extraction from literature with multi-level gene normalization. *PLoS One*, **8**, e55814.
18. Cohen, A.M., Hersh, W.R. (2006) The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *J Biomed Discov Collab*, **1**, 4.
19. Lu, Z., Cohen, K.B., Hunter, L. (2006) Finding GeneRIFs via gene ontology annotations. *Pac Symp Biocomput*, 52-63.
20. Lu, Z. (2007) Text Mining on GeneRIFs. *Computaitonal Bioscience Program*. University of Colorado School of Medicine, Aurora, USA, Vol. Ph.D. thesis.
21. Huang, M., Neveol, A., Lu, Z. (2011) Recommending MeSH terms for annotating biomedical articles. *J Am Med Inform Assoc*, **18**, 660-667.
22. Neveol, A., Shooshan, S.E., Humphrey, S.M.*, et al.* (2009) A recent advance in the automatic indexing of the biomedical literature. *Journal of biomedical informatics*, **42**, 814-823.
23. Vasuki, V., Cohen, T. (2010) Reflective random indexing for semi-automatic indexing of the biomedical literature. *Journal of biomedical informatics*, **43**, 694-700.
24. Huang, M., Lu, Z. (2010) Learning to annotate scientific publications. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, Beijing, China, pp. 463-471.
25. Auken, K.V., Schaeffer, M.L., McQuilton, P.*, et al.* (2013) Corpus Construction for the BioCreative IV GO Task. *Proceedings of the BioCreative IV workshop*, Bethesda, USA.
26. Eisner, R., Poulin, B., Szafron, D.*, et al.* (2005) Improving protein function prediction using the hierarchical structure of the Gene Ontology. *Proceedings of 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*.
27. Kiritchenko, S., Matwin, S., Famili, A.F. (2005) Functional annotation of genes using hierarchical text categorization. *Proceedings of the BioLINK SIG: Linking Literature, Information and Knowledge for Biology*.
28. Comeau, D.C., Islamaj Dogan, R., Ciccarese, P.*, et al.* (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database : the journal of biological databases and curation*, **2013**, bat064.